Simulation studies and Results for the Paper "A Second-Order Longitudinal Model for Binary Outcomes: Item Response Theory versus Structural Equation Modeling"

Part I: A simulation study

Simulation design

A small-scale simulation study is designed to evaluate the performance of the two estimation methods, FIML, termed "ML" in M*plus* syntax, and WLSMV, in terms of convergence rate and model parameter recovery. The only manipulated factor is the proportion of anchor items, which is set to be either 20% or 40%. Kolen and Brennan (2004) recommended that at least 20% items need to be shared between different test forms to have enough information to link the scale. We think this is the most interesting manipulated factor. This is because having more anchor items should lead to better parameter recovery in the different-anchor design, yet it introduces higher computational burden, due to the larger number of testlet factors involved, in the same-anchor design. We believe this also is the first study to systematically compare the behavior of the two estimation methods in the presence of anchor items and a repeated measure data structure. Examinee sample size is 1000, and the test length is 20. We intentionally choose not to vary these factors because sample size and test length have been studied extensively in the IRT literature (*e.g.*, Wang & Nydick, 2014), and their effects on model recovery is well-documented. In all, we have 2 (proportion of anchor items) × 2 (different vs. same anchor designs) = 4 conditions. Within each condition, both (FI)ML and WLS(MV) methods are used, and 100 replications are performed.

Data generation

Assuming there are four time points, person parameters are generated following three steps. First simulate growth factors from a normal distributions, i.e., π_{0i}~N(0,0.4), π_{1i}~N(0.25,0.01) (Kohli, Hughes, Wang, Zopluoglu, & Davison, 2015). Then, given π_{0i} and π_{1i} for person *i*, simulate examinees' ability at each time point, where the residual variance is fixed at 0.15 (Kohli, et al., 2015). Even though residual variance often is manipulated in growth model literature, we choose to keep it fixed again because its effect is well known and thus less interesting. Second, simulate item parameters a_j~U(1.5, 2.5), and b_j~N(0,1) (Cai, 2010). The loadings on the testlet factor in the same-anchor design are simulated as a_{s_j}~U(0.5, 0.6), resulting in the residual correlation between 0.25 and 0.36 (Serrano, 2010).

Model Fitting in M*plus*

Because the item parameters are generated from the IRT logistic metric, item parameters must be translated to the SEM framework using pre-transformations and conversely, M*plus* parameter estimates must be translated back to the IRT logistic metric for comparison to the true values using the post-transformation equations. Please see the Appendix for a complete example of an M*plus* script that we use.

Because FIML involves integrating out the incidental parameters, which number 4 in the different-anchor design, or 8 (20% condition) or 12 (40% condition) in the same-anchor design, quadratic-based numerical integration likely is very time-consuming. Therefore, for highly complex models of many dimensions, a simulated integral, such as Monte-Carlo integration, is adopted.

Evaluation Criteria

The performance of the two estimation methods under different manipulated conditions are evaluated by the following criteria: (1) convergence rate, defined as the proportion of

converged replications out of 100 replications; and (2) average bias and root mean squared error (RMSE), both of which are computed only for successful, *i.e.*, converged, replications. The average bias for item parameters, *a* and *b*, on the non-anchored, unique, items are computed using the equation, $\frac{1}{J}\sum_{j=1}^{J}(\hat{a}_{jt} - a_{jt})$, where *J* denotes the total number of non-anchored items at time *t*.

Given the average biases computed from successful replications, the median value is reported in the final tables. We choose the median because it is less affected by the extreme values that possibly could exist in certain replications (Wang & Nydick, 2015). Similarly, the RMSE is computed within each replication for item parameters using the formula of $\sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{a}_{jt} - a_{jt})^2}$, and the median value among replications is reported finally.

The median average bias and RMSE are computed in a similar fashion for the person parameters ($\theta_1,\ldots,\theta_4$) as well. As for the intercept, slope, residual variance, and mean and variance of the population ability parameters, the median bias is computed as the median of all bias estimates from successful replications, whereas the RMSE is computed as $\sqrt{\frac{1}{R}\sum_{j=1}^{R}(\hat{\beta}_0^r - \beta_0)^2}$ for mean intercept parameter as an example. Here, *R* denotes the total number of successful replications, and $\hat{\beta}_0^r$ denotes the estimate from the *r*th replication.

Results

Convergence Rate

Figure 1 depicts convergence rates for the different- and same-anchor designs of the two cases, 20% and 40% anchor items, with two estimators. For the different-anchor design, all cases and all estimators have strong rates of convergence, at least 94%. There is no appreciable

difference between WLSMV and FIML methods. For the same-anchor design, it can be seen that the FIML estimator has the higher rate of convergence regardless of percentage of items anchored. Most of the failed replications for the FIML estimator are due to item responses being all the same value, either all correct or all incorrect. A number of replications reach saddle points. M*plus* defines the saddle point as the condition when the first derivative of the log-likelihood is indeed 0 but the second derivative matrix is not positive definite. If numerical integration is used in the evaluation of the likelihood, common when Monte Carlo integration is employed as in our study, the information matrix is estimated with error and sometimes the variance covariance matrices for the latent variables can be nearly singular. For the WLSMV estimator, as the percentage of items anchored increases, meaning the number of parameters to be estimated increases, the rate of convergence drops noticeably. For this estimator, in addition to the issue of item responses having the same value, as with the ML estimator, most replications fail due to what M*plus* described as either non-convergence or a lack of identifiability. No replication producing any of these warning or error messages is included in the computation of results.

Parameter Recovery

Different-anchor design Table 1 presents the recovery of item parameters in the different-anchor design. As clearly shown, the median average bias of the discrimination parameter is relatively high across all time points with the FIML estimator when only 20% of items are anchored, whereas the median RMSE from both FIML and RMSE are rather comparable without any noticeable difference. When 40% of items are anchored, the median average bias of the discrimination parameter from ML also is elevated as compared to WLSMV, but of smaller, more acceptable, magnitude. It is well-documented that *a*-parameters are

relatively harder to recover than *b*-parameters (*e.g.*, Wang, Fan, Chang, & Douglas, 2013); that is why the RMSE of *a* almost double that of *b*. Increasing the proportion of anchor items helps decrease the average bias and RMSE for ML, yet not so much for WLSMV. One possible explanation is that WLSMV relies on the first four moments of raw data, which are less affected by less information, *i.e.*, 20% anchors, whereas ML uses raw response data and thus it benefits from having more information, *i.e.*, 40% anchors.

Table 2 shows the recovery of person parameters which is all quite good. Both median average bias and median RMSE are acceptable. Neither estimation method nor proportion of anchors has much impact on virtually all person parameters. Table 3 presents the recovery of mean and variance of the intercept and slope, as well as residual variance, all of which are directly *estimated* from the model. In addition, we also present the recovery of the population means and variances of θ's at four points, which are *computed* from Eqs. (4) and (5). These values paint out the group growth trajectories. Table 4 shows the median average bias and median RMSE are uniformly small in all cells, implying the successful recovery of these parameters pertaining to growth.

Same-anchor design Even though the same-anchor design introduces extra difficulty for model convergence, for those successfully converged replications, the quality of the parameter estimates is rather comparable to that from the different-anchor design, as reflected in Tables 4 – 6. More specifically, in Table 4, the median average bias for *a*-parameter is somewhat elevated for the FIML estimator when compared to the WLSMV estimator, but not unacceptably so. This effect is more muted for the larger percentage of items anchored. The ML estimator produces uniformly smaller median RMSE than the WLSMV estimator for both *a*- and *b*- parameters, and when there is larger percentage of anchor items, the median RMSE for both methods decreases.

The loadings on the testlet factor also are recovered well for both methods and both manipulated scenarios. One interesting observation for the testlet loadings is that its median average bias and median RMSE stay constant over four time points under FIML, whereas the median average bias and median RMSE vary slightly under WLSMV. This is because, as alluded to earlier, the true discrimination parameters on the testlet factors are fixed as constants over time as are the estimated values using FIML. When the WLSMV method is called upon in M*plus*, even though we constrain the loadings on the testlet factor to be steady over time, the post-transformation involves a time-variant factor variance term, namely, $\widehat{Var}(\eta_t)$. Thus, the estimate of the testlet slope under the WLSMV estimator, when translated back to the IRT framework, varies somewhat across time points. It can be seen that this effect is slight.

The person parameters, growth parameters, mean and variance of the population abilities all recover quite well for the same-anchor design. In summary, the same-anchor design has good parameter recovery performance when the replication converges. This design has more difficulty with convergence than the different-anchor design. The WLSMV estimator tends to generate slightly lower median average bias, but has higher median RMSE than the ML estimator. Additionally, it is much faster, seconds vs. 10+ hours for each replication, than the ML estimator in the same-anchor design.

Table 1. Recovery of item parameters in the different-anchor design

Percent Items Anchored	Metric	Estimator	Time 1		Time 2		Time 3		Time 4	
			a	b	a	b	a	b	a	b
20%	Median Average Bias	FIML	-0.1149	-0.0033	0.127	0.008	0.1344	0.0173	0.1089	0.0208
		WLSMV	-0.0305	0.0034	-0.012	0.0038	0.0233	0.0122	0.0323	0.0119
	Median RMSE	FIML	0.2459	0.0978	0.2595	0.1053	0.2811	0.1263	0.2635	0.1284
		WLSMV	0.2455	0.1198	0.2498	0.11	0.2696	0.1415	0.2688	0.1407
40%	Median Average Bias	FIML	0.0675	0.0017	0.0406	0.006	0.0507	0.0109	0.0526	0.0091
		WLSMV	-0.0269	0.002	-0.0274	-0.0029	-0.0105	-0.0055	-0.0464	-0.0095
	Median RMSE	FIML	0.2072	0.083	0.2045	0.0844	0.218	0.0888	0.2095	0.1076
		WLSMV	0.2423	0.1036	0.2469	0.1146	0.2603	0.1114	0.2704	0.1393

Table 2. Recovery of person parameters in the different-anchor design

Percent Items Anchored	Estimator	Median Average Bias				Median RMSE			
		θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
20%	FIML	0.0032	3.00E-04	0.0025	0.0011	0.2824	0.2797	0.2891	0.3074
	WLSMV	0.0035	-1.00E-04	-0.0134	-0.0173	0.2694	0.2678	0.2766	0.2955
40%	FIML	-6.00E-04	-4.00E-04	0.005	0.0038	0.276	0.2761	0.2818	0.296
	WLSMV	0.0014	-0.0067	-0.0084	-0.01	0.2686	0.2692	0.2758	0.2915

Table 3. Recovery of intercept, slope, residual variance, mean, and variance of the population ability in the different-anchor design

Percent Items Anchored	Metric	Estimator	Estimated					Calculated							
			β_0	β_1	υ_{0i}	υ_{1i}	σ_ε^2	μ_{θ_1}	μ_{θ_2}	μ_{θ_3}	μ_{θ_4}	$\sigma_{\theta_1}^2$	$\sigma_{\theta_2}^2$	$\sigma_{\theta_3}^2$	$\sigma_{\theta_4}^2$
20%	Median Bias	FIML	5.00E-04	0.0045	0.044	0.002	-0.034	-5.00E-04	-0.0035	-0.003	-0.011	-0.075	-0.0775	-0.0825	-0.0925
		WLSMV	-0.002	0.006	0.005	0.001	-0.006	0.002	-0.002	-0.0085	-0.011	-0.0105	-0.012	-0.017	-0.021
	RMSE	FIML	0.0367	0.0182	0.0528	0.0036	0.0349	0.0367	0.0294	0.0322	0.0432	0.0853	0.0867	0.0915	0.1015
		WLSMV	0.0322	0.0144	0.0468	0.008	0.0284	0.0322	0.0276	0.0301	0.0382	0.0639	0.0572	0.0404	0.0383
40%	Median Bias	FIML	0	0.001	0.017	0.001	-0.02	0	-0.002	0.002	0.002	-0.036	-0.036	-0.04	-0.042
		WLSMV	0	0.004	0.0055	0.001	-0.005	0	-0.001	0	-0.0035	-0.009	-0.0105	-0.0135	-0.0185
	RMSE	FIML	0.0249	0.0102	0.0291	0.0032	0.0233	0.0249	0.0213	0.0224	0.0274	0.0459	0.0465	0.0494	0.0575
		WLSMV	0.025	0.0098	0.0186	0.0031	0.01	0.025	0.0218	0.0229	0.0277	0.0217	0.021	0.0216	0.0296

Table 4. Recovery of item parameters in the same-anchor design

Percent Items Anchored	Metric	Estimator	Time 1			Time 2			Time 3			Time 4		
			a	b	Testlet Slope	a	b	Testlet Slope	a	b	Testlet Slope	a	b	Testlet Slope
20%	Median Average Bias	**FIML**	0.0529	-0.0066	-0.0572	0.0581	0.0011	-0.0572	0.0689	0.0012	-0.0572	0.0628	0.0113	-0.0572
		WLSMV	0.0091	-0.0063	-0.0427	0.0278	-0.0137	-0.0418	-0.0023	-0.0296	-0.0391	0.0219	-0.0288	-0.0304
	Median RMSE	**FIML**	0.2226	0.0936	0.1269	0.22	0.0877	0.1269	0.2371	0.1144	0.1269	0.2335	0.1279	0.1269
		WLSMV	0.2638	0.1279	0.1276	0.2571	0.1003	0.1274	0.2817	0.1602	0.1273	0.2912	0.1413	0.1257
40%	Median Average Bias	**FIML**	0.0364	-0.0016	-0.0292	0.0264	-0.0032	-0.0292	0.0295	0.0022	-0.0292	0.0325	-0.0056	-0.0292
		WLSMV	0.0179	-0.0059	-0.049	0.0256	-0.0041	-0.0477	0.0094	-0.0133	-0.0432	0.0102	-0.0262	-0.0334
	Median RMSE	**FIML**	0.2031	0.0835	0.1203	0.2023	0.0758	0.1203	0.2083	0.0938	0.1203	0.2118	0.1039	0.1203
		WLSMV	0.2395	0.0853	0.1465	0.2337	0.0875	0.1454	0.2589	0.1375	0.1439	0.2742	0.1363	0.1443

Table 5. Recovery of person parameters in the same-anchor design

Percent Items Anchored	Estimator	Median Average Bias				Median RMSE			
		θ_1	θ_2	θ_3	θ_4	θ_1	θ_2	θ_3	θ_4
20%	FIML	-2.00E-04	-0.0017	-0.0014	-0.0037	0.2748	0.2699	0.2801	0.296
	WLSMV	2.00E-04	-0.0082	-0.023	-0.033	0.2735	0.271	0.2827	0.2984
40%	FIML	-0.0013	-0.0036	-0.0026	-0.006	0.2729	0.2709	0.2772	0.2946
	WLSMV	-0.0061	-0.0126	-0.0207	-0.0371	0.2774	0.2734	0.2803	0.298

Table 6. Recovery of intercept, slope, residual variance, mean, and variance of the population ability in the same-anchor design

Percent Items Anchored	Metric	Estimator	Estimated					Calculated							
			β_0	β_1	υ_{0i}	υ_{1i}	σ^2	μ_{θ_1}	μ_{θ_2}	μ_{θ_3}	μ_{θ_4}	$\sigma^2_{\theta_1}$	$\sigma^2_{\theta_2}$	$\sigma^2_{\theta_3}$	$\sigma^2_{\theta_4}$
20%	Median Bias	FIML	-0.001	0	-0.014	0	-0.013	-0.001	0.001	0.002	0.004	-0.03	-0.03	-0.027	-0.028
		WLSMV	0.001	-0.005	-0.03	-0.001	-0.014	0.001	-0.0055	-0.0125	-0.0165	-0.0435	-0.044	-0.049	-0.052
	RMSE	FIML	0.0315	0.0146	0.0344	0.0032	0.0178	0.0315	0.0259	0.0278	0.0361	0.0488	0.0488	0.0501	0.056
		WLSMV	0.0304	0.0141	0.0371	0.003	0.0167	0.0304	0.0274	0.0314	0.0402	0.0514	0.0521	0.0554	0.0631
40%	Median Bias	FIML	0	-0.001	-0.005	0.001	-0.007	0	-0.001	0.003	0.002	-0.014	-0.013	-0.012	-0.008
		WLSMV	5.00E-04	-0.005	-0.028	-5.00E-04	-0.0145	5.00E-04	-0.009	-0.015	0.0225	-0.0405	-0.043	-0.045	-0.051
	RMSE	FIML	0.0242	0.0103	0.0251	0.0028	0.0121	0.0242	0.024	0.028	0.0347	0.0317	0.0314	0.0316	0.0365
		WLSMV	0.0239	0.0108	0.0331	0.0026	0.0164	0.0239	0.0246	0.0294	0.0371	0.0463	0.0466	0.0484	0.0538

Part II: A pilot study based on real data (sensitivity analysis)

Protocol

Based on the NELS blueprint, 13 items were administered all three times and eleven items were administered on two occasions. This gives rise to three modeling scenarios:

a. All 24 repeated items are represented by 24 corresponding testlets (Model 1)
b. The 13 testlets administered all three times are represented by 13 testlets (Model 2)
c. No testlets are included in the M*plus* model (Model 3)

In all three cases, all 24 repeated item parameter values are anchored in the M*plus* input code to the estimates reported in the NELS:88 report, regardless of whether a corresponding testlet is included.

Results

(1) Model 1

Table 1: Recovery of Item Parameters

Testlet Loading	Metric	Estimator	Time 1			Time 2			Time 3		
			a	b	Testlet Slope	a	b	Testlet Slope	a	b	Testlet Slope
0.1	Average Bias	ML	0.149	-0.0469	0.0289	0.0421	0.2813	0.0318	-0.0454	1.3428	0.0181
		WLSMV	0.1489	-0.2501	-0.0378	0.0441	0.2067	-0.0198	-0.0613	1.477	-0.0433
	RMSE	ML	0.191	0.3777	0.1681	0.0421	0.2813	0.1912	0.173	2.0702	0.1691
		WLSMV	0.169	0.5029	0.2483	0.0441	0.2067	0.2837	0.1845	2.4087	0.2888
0.5	Average Bias	ML	0.143	-0.1727	-0.2994	0.1156	0.0389	-0.2953	-0.0426	1.3989	-0.3037
		WLSMV	0.1229	-0.319	-0.3337	0.1371	0.0265	-0.3902	-0.0546	1.6149	-0.3381
	RMSE	ML	0.1599	0.396	0.3535	0.1156	0.1147	0.3561	0.1914	2.3336	0.3638
		WLSMV	0.1431	0.4964	0.4646	0.1371	0.1167	0.5105	0.2085	2.8601	0.4862

Table 2: Recovery of Person Parameters

Testlet Loading	Estimator	Median Bias			RMSE		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
0.1	ML	0.1632	0.09	0.131	0.5687	0.5619	0.5842
	WLSMV	0.0612	0.0396	0.0627	0.5182	0.4994	0.5102
0.5	ML	0.1586	0.0919	0.1209	0.578	0.5509	0.575
	WLSMV	0.0567	0.0306	0.0565	0.5123	0.496	0.5079

Table 3: Recovery of Intercept, Slope, Mean, and Variance of the Population Ability

Testlet Loading	Metric	Estimator	Estimated					Calculated					
			β_0	β_1	v_{0i}	v_{1i}	σ^2	μ_{θ_1}	μ_{θ_2}	μ_{θ_3}	$\sigma^2_{\theta_1}$	$\sigma^2_{\theta_2}$	$\sigma^2_{\theta_3}$
0.1	Average Bias	ML	0.142	-0.016	0	0	1.953	0.142	0.147	0.14	0.953	0.953	1.052
		WLSMV	0.06	-0.004	0.011	0	1.8615	0.06	0.048	0.0445	0.872	0.872	0.922
	RMSE	ML	0.1682	0.025	0	0.0087	1.9892	0.1682	0.1421	0.1303	0.9926	1.0043	1.0457
		WLSMV	0.057	0.0135	0.0357	0.006	1.8463	0.057	0.0528	0.0612	0.8686	0.8804	0.9183
0.5	Average Bias	ML	0.1425	-0.0075	0	0	1.9275	0.1425	0.127	0.1165	0.9465	0.9465	0.955
		WLSMV	0.052	0.001	0.001	0	1.795	0.052	0.04	0.028	0.805	0.814	0.829
	RMSE	ML	0.146	0.015	0.0437	0.0042	1.9014	0.146	0.131	0.1215	0.9188	0.9223	0.935
		WLSMV	0.0578	0.0133	0.0339	0.0024	1.7828	0.0578	0.0527	0.0605	0.8023	0.8079	0.8253

(2) Model 2

Table 1: Recovery of Item Parameters

Testlet Loading	Metric	Estimator	Time 1			Time 2			Time 3		
			a	b	Testlet Slope	a	b	Testlet Slope	a	b	Testlet Slope
0.1	Average Bias	ML	0.1613	-0.094	-0.0152	0.0571	0.2207	-0.0152	-0.0649	1.3913	-0.0152
		WLSMV	0.1313	-0.3208	-0.0773	0.0116	0.2623	-0.0773	-0.0477	2.18	-0.0773
	RMSE	ML	0.1871	0.3339	0.1314	0.0571	0.2207	0.1314	0.1798	2.0536	0.1314
		WLSMV	0.1555	0.5539	0.2404	0.0116	0.2623	0.2405	0.2389	4.1146	0.2405
0.5	Average Bias	ML	0.1469	-0.1586	-0.3319	0.1011	0.0609	-0.3319	-0.0389	1.309	-0.3319
		WLSMV	0.1412	-0.3131	-0.4063	0.1187	0.0401	-0.4063	-0.0472	1.6302	-0.4063
	RMSE	ML	0.1653	0.3893	0.3731	0.1011	0.1145	0.3731	0.1815	2.1911	0.3731
		WLSMV	0.1546	0.4842	0.5394	0.1187	0.1043	0.5394	0.2008	2.9521	0.5395

Table 2: Recovery of Person Parameters

Testlet Loading	Estimator	Median Bias			RMSE		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
0.1	ML	0.1578	0.0919	0.1324	0.5816	0.5553	0.5793
	WLSMV	0.0676	0.0452	0.0645	0.5186	0.505	0.5136
0.5	ML	0.1479	0.0854	0.1167	0.5688	0.5472	0.5662
	WLSMV	0.0478	0.0266	0.0576	0.5095	0.4991	0.5103

Table 3: Recovery of Intercept, Slope, Mean, and Variance of the Population Ability

Testlet Loading	Metric	Estimator	Estimated					Calculated					
			β_0	β_1	v_{0i}	v_{1i}	σ^2	μ_{θ_1}	μ_{θ_2}	μ_{θ_3}	$\sigma^2_{\theta_1}$	$\sigma^2_{\theta_2}$	$\sigma^2_{\theta_3}$
0.1	Average Bias	ML	0.138	-0.003	0	0	1.951	0.138	0.13	0.142	0.955	0.955	1.009
		WLSMV	0.062	0.002	0	0	1.87	0.062	0.07	0.085	0.87	0.871	0.9
	RMSE	ML	0.1398	0.0215	0.0254	0.0069	1.9376	0.1398	0.1295	0.1333	0.9508	0.9651	1.0112
		WLSMV	0.0747	0.0135	0.0309	0.0058	1.8567	0.0747	0.0666	0.069	0.8705	0.8801	0.9114
0.5	Average Bias	ML	0.13	-0.005	0	0	1.861	0.13	0.115	0.1	0.899	0.906	0.917
		WLSMV	0.052	5.00E-04	0	0	1.781	0.052	0.05	0.0455	0.8005	0.805	0.823
	RMSE	ML	0.132	0.014	0.0442	0.0044	1.8581	0.132	0.1178	0.109	0.8807	0.8876	0.9091
		WLSMV	0.0572	0.0119	0.0483	0.0033	1.7725	0.0572	0.0518	0.057	0.8005	0.8063	0.8248

(3) Model 3

Table 1: Recovery of Item Parameters

Testlet Loading	Metric	Estimator	Time 1			Time 2			Time 3		
			a	b	Testlet Slope	a	b	Testlet Slope	a	b	Testlet Slope
0.1	Average Bias	ML	0.1488	-0.1871	N/A	0.0521	0.214	N/A	-0.0745	1.4036	N/A
		WLSMV	0.1025	-0.3748	N/A	0.0893	0.1278	N/A	-0.0619	1.5127	N/A
	RMSE	ML	0.1877	0.3838	N/A	0.0521	0.2183	N/A	0.1811	2.2469	N/A
		WLSMV	0.1242	0.5918	N/A	0.0893	0.1389	N/A	0.1933	2.5957	N/A
0.5	Average Bias	ML	0.1686	-0.0982	N/A	0.0561	0.1862	N/A	-0.0624	1.4559	N/A
		WLSMV	0.1121	-0.3047	N/A	0.1138	0.1039	N/A	-0.0569	1.5425	N/A
	RMSE	ML	0.2005	0.3477	N/A	0.0561	0.1862	N/A	0.1884	2.0794	N/A
		WLSMV	0.1357	0.507	N/A	0.1138	0.1039	N/A	0.2074	2.4177	N/A

Table 2: Recovery of Person Parameters

Testlet Loading	Estimator	Median Bias			RMSE		
		θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
0.1	ML	0.1384	0.1614	0.1465	0.5518	0.5467	0.5409
	WLSMV	0.0855	0.0804	0.0828	0.4457	0.4441	0.4487
0.5	ML	0.1066	0.1413	0.1245	0.5198	0.5136	0.5093
	WLSMV	0.0662	0.0742	0.0724	0.4361	0.4358	0.4337

Table 3: Recovery of Intercept, Slope, Mean, and Variance of the Population Ability

Testlet Loading	Metric	Estimator	Estimated					Calculated					
			β_0	β_1	v_{0i}	v_{1i}	σ^2	μ_{θ_1}	μ_{θ_2}	μ_{θ_3}	$\sigma^2_{\theta_1}$	$\sigma^2_{\theta_2}$	$\sigma^2_{\theta_3}$
0.1	Average Bias	ML	0.168	-0.0065	2.059	0	0	0.168	0.1635	0.1385	1.0595	1.06	1.064
		WLSMV	0.07	-0.002	1.881	0	0	0.07	0.066	0.054	0.883	0.892	0.9
	RMSE	ML	0.1816	0.0113	2.1077	2.00E-04	0.0016	0.1816	0.1669	0.1543	1.1133	1.1138	1.1152
		WLSMV	0.0803	0.0088	1.8899	0.0017	0.009	0.0803	0.077	0.0776	0.895	0.898	0.9069
0.5	Average Bias	ML	0.136	-0.008	1.968	0	0.002	0.136	0.122	0.107	0.968	0.968	0.97
		WLSMV	0.041	-0.002	1.784	0	0	0.041	0.041	0.036	0.788	0.788	0.793
	RMSE	ML	0.1448	0.0108	1.974	0	0.0018	0.1448	0.1284	0.1138	0.9776	0.978	0.9788
		WLSMV	0.0613	0.0075	1.7939	8.00E-04	0.0062	0.0613	0.0542	0.0508	0.7974	0.7985	0.802